Possible lines of work

Feb 2017

Language-based video understanding

- methods that leverage both textual and visual modalities to infer or predict sequence data.
  - Video captioning
  - Video story generation
  - Object tracking
  - Video-text alignment
  - Description-based action detection
  - ...

Language-based video Tracking

- improve CVPR’17 submission
  - Initializing tracker with language-specification model is error pruning and cause error propagation.
  - In many videos, a single word is enough for description i.e. language-based tracking does not make sense. What is the performance of method with just one word as description?
  - The dataset is a challenge. Do we need more supervision during training? For example by not only just describing first frame of the video but also next frames during training.
  - Improving joint tracker by investigating different trackers?
  - Finding error sources is important. Better visualization helps.

Language-based event detection

Alignment -> aligning video to book, video to synopses, video to chapters in the book.
- shots and sentences are available both for training and testing

Detection - similar picture to CVPR'17 submission but for the purpose of temporally locating the actions (and possibly spatially if necessary).

- Temporal sections of the videos are easier to describe
  - E.g. In this video, a man first called somebody then wore his clothes, went out of his house and then met a guy.
- shots and sentences are available just in training time.
- Horizon: its extension to online event detection? Weakly supervised learning where just order of the actions is provided during training.

Textual question / visual answering (related to retrieval?)

- Given long video X and a query “what did appear after kissing?”.
- OR: given a set of video shots and the query Q, retrieve/rank video shots.

Datasets with textual annotations

- MovieQA: Understanding Stories in Movies through Question-Answering
  - # of movies 140
  - Multi source of information is provided (scripts, plots, video, subtitle, DVS)
  - No temporal annotation (x)
- Large Scale Movie Description Challenge (LSMDC)
  - 202 movies
  - Text extracted from Audio Descriptions (AD) of movies.
  - 118k sentences, temporally aligned
- TACoS
  - Temporal annotations, multi-level description

Detailed: A man took a cutting board and knife from the drawer. He took out an orange from the refrigerator. Then, he took a knife from the drawer. He juiced one half of the orange. Next, he opened the refrigerator. He cut the orange with the knife. The man threw away the skin. He got a glass from the cabinet. Then, he poured the juice into the glass. Finally, he placed the orange in the sink.
Short: A man juiced the orange. Next, he cut the orange in half. Finally, he poured the juice into a glass.
One sentence: A man juiced the orange.

Agent-based object detection

Reinforcement learning for detection: A fundamentally different strategy.

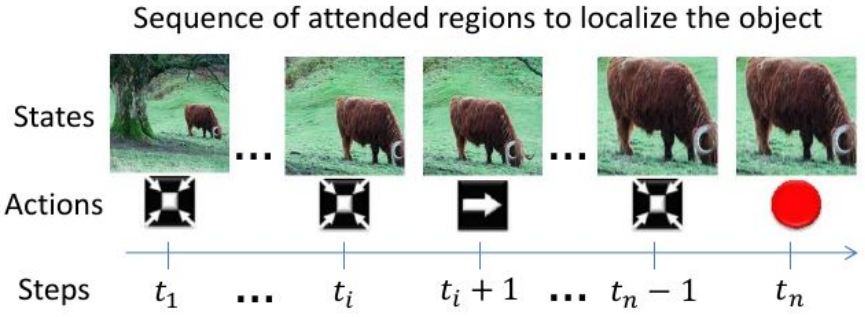


Figure 1. A sequence of actions taken by the proposed algorithm to localize a cow. The algorithm attends regions and decides how to transform the bounding box to progressively localize the object.

- One agent per category
- 200 steps to localize each box
- Not yet very mature =>
- Localized boxes are used as object proposals

Agent-based action detection

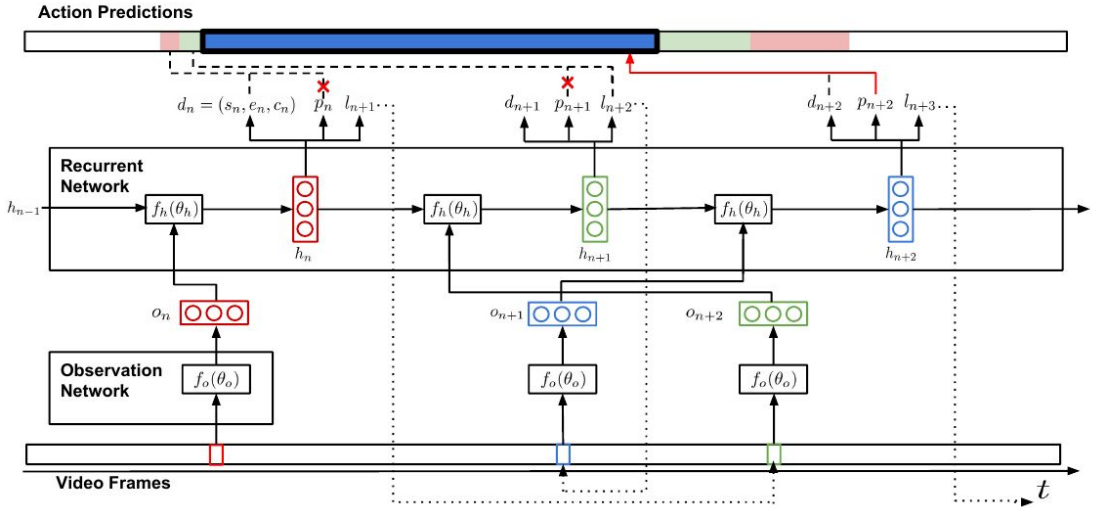


Figure 2: The input to the model is a sequence of video frames, and the output is a set of action predictions. We illustrate an example of a forward pass. At timestep n, the agent observes the red video frame and produces candidate detection d_n. d_n is shown for reference on the timeline of action predictions, however prediction indicator output p_n suppresses it from being emitted into the prediction set (indicated by the red cross). Observation location output l_{n+1} signals to observe the the green video frame at the next timestep. The process repeats, and here again p_{n+1} suppresses d_{n+1} from being emitted. l_{n+2} signals to now go backwards in the video to observe the blue frame. At timestep $n+2$, the action hypothesis is sufficiently refined, and the agent uses prediction indicator p_{n+2} to emit d_{n+2} into the prediction set (red arrow). The agent then continues proceeding through the video.

Natural language detection

- Natural Language Object Retrieval
  - In contrast to classic object detection, that there is one model for each object category, in this work one model is learned for all the sentences => class-agnostic
  - If we train the model with one word instead of one sentence, can such architecture be considered as an object proposal generator?

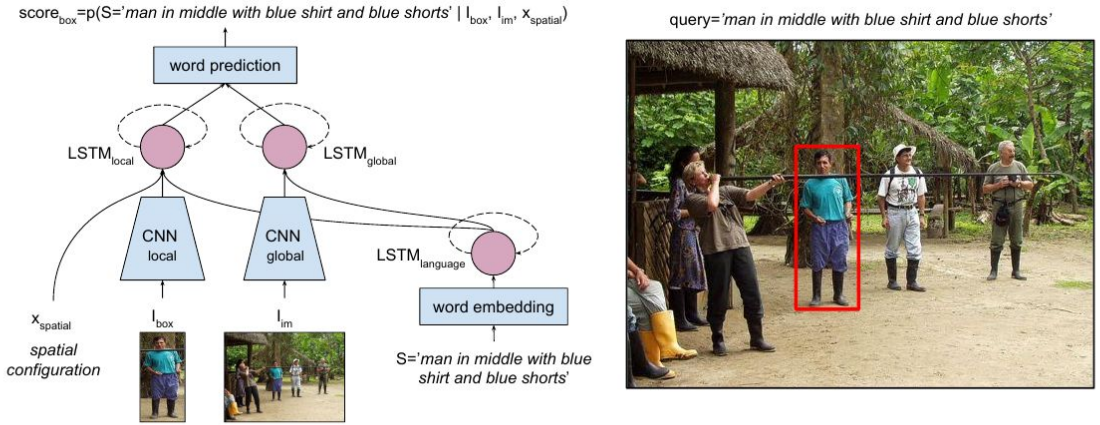


Figure 2. Our Spatial Context Recurrent ConvNet (SCRC) for natural language object retrieval. The recurrent network in our model contains three LSTM units. Two CNN's are used to extract local image descriptors and global scene-level contextual feature respectively. Parameters in word embedding, word prediction and three LSTM units are initialized by pretraining on image captioning dataset.

Language-based event retrieval

- The problem with natural language object retrieval is that the query is specific for each image and can not be generalized to other images
  - E.g. “man in middle with blue shirt and blue shorts’.
- Moving to video, the actions are temporally in a more structured order
  - E.g. “singing while blowing out the candles”.
  - Spatial location of object in videos are less important. (yes?)
- Problem: given a description to an agent, it finds video snippets corresponding to the description.
  - A one-to-many mapping from one description to many snippets in a long video.
  - Take into account order of actions/objects in the description while finding snippets.

Datasets

Table 1. Comparison of Charades with other video datasets.

	Actions per video	Classes	Labelled instances	Total videos	Origin	Type	Temporal localization
Charades v1.0	6.8	157	67K	10K	267 Homes	Daily Activities	Yes
ActivityNet [3]	1.4	203	39K	28K	YouTube	Human Activities	Yes
UCF101 [8]	1	101	13K	13K	YouTube	Sports	No
HMDB51 [7]	1	51	7K	7K	YouTube/Movies	Movies	No
THUMOS'15 [5]	1-2	101	21K+	24K	YouTube	Sports	Yes
Sports 1M [6]	1	487	1.1M	1.1M	YouTube	Sports	No
MPII-Cooking [14]	46	78	13K	273	30 In-house actors	Cooking	Yes
ADL [25]	22	32	436	20	20 Volunteers	Ego-centric	Yes
MPII-MD [11]	Captions	Captions	68K	94	Movies	Movies	No

Action labels vs. description

Annotated Actions: (gray if not active)

Holding a cup/glass/bottle of something
Taking a towel/s from somewhere
Holding a towel/s
Drinking from a cup/glass/bottle
Putting a cup/glass/bottle somewhere
Opening a closet/cabinet
Opening a door
Grasping onto a doorknob
Putting a towel/s somewhere
Putting clothes somewhere
Taking a broom from somewhere
Throwing a towel/s somewhere
Closing a door
Closing a closet/cabinet
Holding a vacuum

Video 4 of 50: (3x Speed)

Annotated Objects:

Broom, Closet/cabinet, Clothes, Door, Glass, Towel, Vacuum

Script:

A person walks towards pantry drinking some water and holding a towel over one arm. The person, throws towel in pantry and takes out vacuum.

0:43 / 8:09

Session #2

27-02-2017

Idea1: learning to describe tracks

- **intro:** there are a few works on describing referring objects in images
  - Given an image and a bounding box, the task is to describe that particular object relative to other objects: “a man with blue shirt on the right side of a TV”
- **idea:** to describe a tracking object: the man goes away from the tree and is approaching to the building”.
- **extension of NLP-tracker:** this task can be considered as an extension of NLP-tracker where the output is both visual tracking and textual tracking. they may also can help each other.

Idea2: action-based NLP-tracker

- **intro:** in current NLP-tracker, an assumption is that the object to be tracked is presented in the first frame.
  - No temporal detection mechanism.
- **idea:** action-based NLP-tracking is an extension of NLP-tracker: a conditional tracker that starts to track when the desired action started to happen e.g. “track the man with blue shirt when he starts to run/he is close to the street”.
-

Idea3: action detection using textual modality

- **intro:** So far action detection is done on visual data.
  - Given a video, the task is to learn an action model to detect the corresponding action in a long video.
- **Idea:** to leverage text to improve the action detection.
  - Training: a tuple of <videos,annotations,text>. Testing: given just a video, detect the action.
- **Motivation:** in image domain, there is not a direct/straight-forward relation between image detection and image description i.e. generated description can not be easily used for the task of image detection. in contrast, for videos, descriptions are along temporal domain i.e. if we describe a video properly, it is likely to detect an action of interest within the description.

Idea3: action detection using textual modality

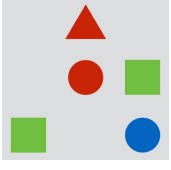
- **Introduction:** Learning image/video description models is getting very popular in recent years. They try to generate a text from a sequence of images.
  - Training: a tuple of <videos,text>. Testing: given a video, describe it.
- **Naive idea:** given a video, a captioning model first generate a text for the given video, then by inspecting the text, one can localize action of interest in the video.
- **A bit nicer:** not to generate human-readable text for a video. Instead to benefit from internal, hidden information in a video captioning model to detect start and end of an action.
  - It leads to design an architecture, combining architectures from the captioning models and the action detection models.

Idea4*: learn to ask questions (self-awareness)

- **intro:** in VQA task, the aim is to train models to answer to the questions
  - In some cases, computers are very uncertain about their decisions
- **idea:** the goal is to train a model that ask questions about what it is not sure about.
  - It can be in an interactive way. However, the evaluation is not easy.
- **Application:** It can be used in interactive captioning systems where computer ask about its uncertainty to generate more precise captions.

Next next step: higher-level understanding

- Visual recognition including classification, detection, tracking, segmentation,... is not an AI-complete problem. They are mostly tackled in a pattern recognition framework.
- Open-world recognition is one-step further.
  - There is not a limited number of classes/categories.
  - Text comes into the game.
- visual reasoning is another step for facing a more general AI problem.
  - Here, we not only need to ground objects and their appearance (like shape and color), but also learn their composition.
  - Recently, it is getting more attention.
  - To me, the first step is to generate a grammar (semantic parser) for describing the scene instead of using a text expression.



Random thoughts

- Any information in feature “velocity” of a sequensial data? E.g. [CNN(fy)-CNN(fx)]/(y-x)?
  - Velocity of features in different layers: what does it say
  - How the information can be extracted?
  - Discovering object based on fired-neurons?